Relationship between Mean and Median

平均値と中央値の関係

Yuuichi KAWAGUCHI
川　口　雄　一

With regard to the relationship between the two averages, mean and median, if the distribution of a sample is symmetric, then both are the same. If however the distribution is skewed, the mean is affected by the tail and outliers much more than the median. The purpose of this paper is to show a counter-example to this principle. Two examples will be presented. The first one satisfies the principle, and the second is a counter-example to it. It is important for teachers to show beginners the existence of a counter-example.

二つの代表値である平均値と中央値の関係の一般原則として、標本の分布は左右対象であればこれら二つは一致する。もしも歪んでいれば、平均値の方が中央値よりも強く「スソ」の影響を受ける。本稿の目的は、この一般原則に対する反例を示すことにある。そのために、原則が成り立つ例と、成り立たない例の二つを示す。統計学を学び始めた学生に対して反例の存在を示すことは教師の重要な役割である。

Key words: average（表値）
mean（平均値）
median（中央値）
skewness（歪度）

天使大学 看護栄養学部 教養教育科
（2009年11月２日受稿、2010年３月３日 審査終了受理）

Ⅰ. Purpose

There are three types of central tendencies (*i.e.*, averages): mean, median, and mode. If a distribution is unimodal, then the following principles are believed to hold in many cases (*e.g.*, [1][2]).

1. If the distribution is symmetric, then the three averages are equal.
2. If the distribution is asymmetric (i.e., skewed), then the mean is more distant from the mode than the median.

In other words, means are affected by tails and outliers much more than medians. Most samples, except for artificial ones, are skewed to a greater or lesser degree. A situation where principle 1 holds is rare. As such, this paper focuses on principle 2. The aim of this paper is to show an affirmative example of and a counter-example to principle 2.

Ⅱ. Affirmative Example

A Sample from a book [3] is shown in Table 1. Each datum denotes the weight of a newborn baby in 1988.

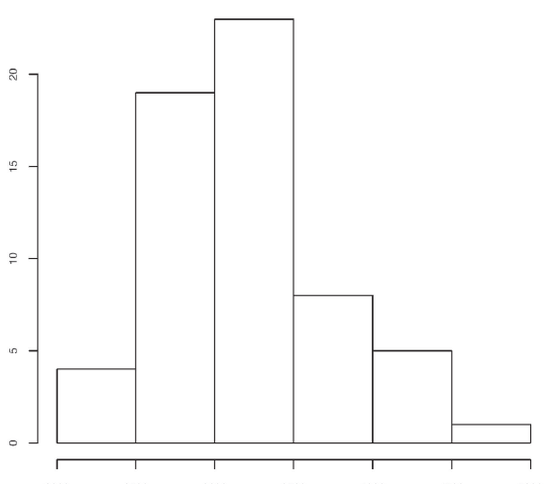

Figure 1. Weight (1988) - (A)

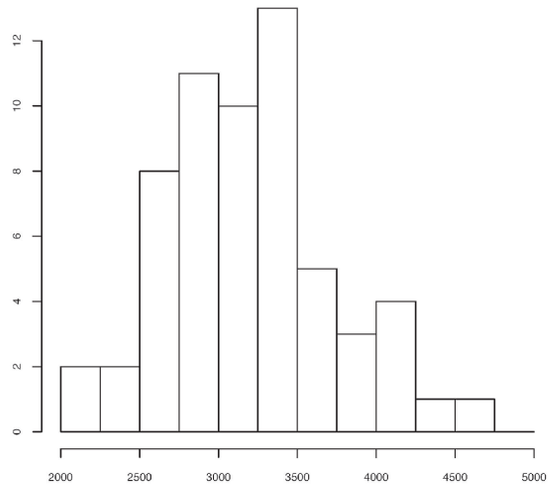

Figure 2. Weight (1988) - (B)

Two histograms are shown in Figures 1 and 2 (one each). The class interval is 500 in Figure 1, and 250 in Figure 2.

The averages, skewness*1, and standard deviation*2 are as follows*3:

size: $n = 60$
mean: $\bar{x} = 3179$
median: 3150
skewness: 0.35
std. dev.: $s = 539$

Skewness is defined as

$$\text{skewness} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n \cdot s^3},$$

where each x_i $(i = 1, \ldots, n)$ is a datum.

Because the skewness is positive, the distribution of the sample is skewed and the right tail is longer than the left one. In this case, according to the principle 2, the median should be lower than the mean, and in fact, the median is lower than the mean.

Thus, this example supports principle 2.

＊ 1 Those are calculated using R [6][5].
＊ 2 The standard deviation function in R (**sd**) uses n-1 as the divisor. Another function that uses n is programmed. It is shown in the appendix.
＊ 3 The precision of those items, except for size and skewness, are as in the original data.

Table 1. Sample of the weights (in *g*) of newborn babies in 1988

3470	2550	2920	2530	3280	2840	2520	3350	3610	3430
3020	3320	2790	3050	3620	3260	3320	3800	2640	3360
3320	4100	2720	4050	3850	3380	3040	2710	4150	3200
4120	2780	3220	2780	2490	2950	2580	2020	3010	2010
2800	3760	4480	2990	3700	2960	2320	3060	3200	3380
3100	2840	2990	3100	3530	3270	2600	3640	3300	4570

Ⅲ. Counter-Example

Another sample from the same book is shown in Table 2. Here, too each datum denotes the weight of a newborn baby in 1978.

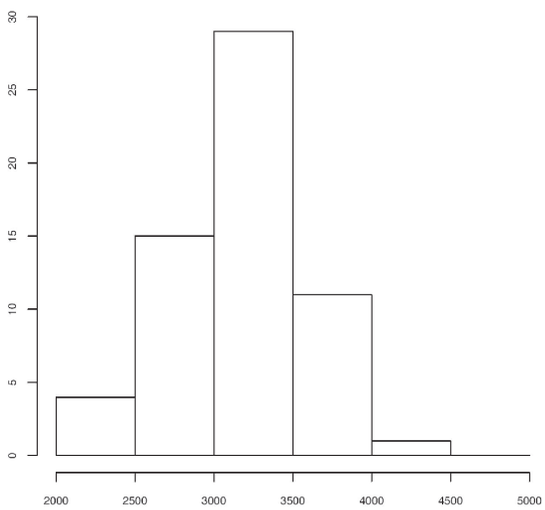

Figure 3. Weight (1978) - (A)

Two histograms are shown in Figures 3 and 4. The class interval is 500 in Figure 3, and 250 in Figure 4.

The averages, skewness, and standard deviation are as follows:

size: $n = 60$
mean: $\bar{x} = 3216$
median: 3200
skewness: -0.15
std. dev.: $s = 409$

Because skewness is negative, the distribution of the sample is skewed and the left tail is longer than the right one. In this case, according to principle 2, the mean should be lower than the median, but in fact, the mean is higher than the median. Thus, principle 2 does not hold.

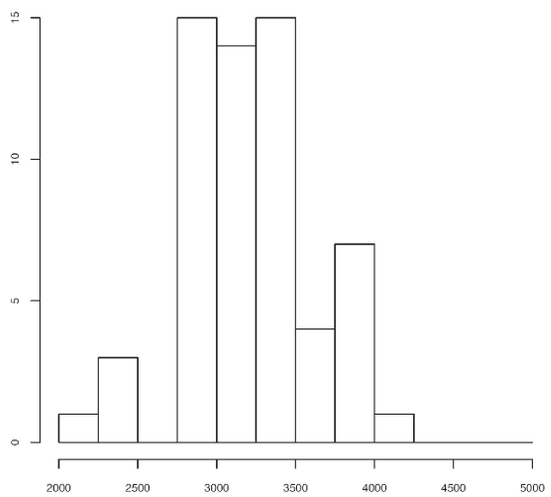

Figure 4. Weight (1978) - (B)

This is a counter-example to principle 2.

Ⅳ. Discussion

The distribution of the sample shown in Table 1 is unimodal. This is confirmed by those histograms shown in Figures 1 and 2. The distribution of the sample shown in Table 2 also seems to be unimodal as can be seen from Figure 3. However, in the histogram shown in Figure 4, there are two peaks. One is at class mark 3,125, and the other is at class mark 3,375. Thus, it is a bimodal distribution and this is the reason why principle 2 does not hold.

A book [4] shows another measure of

Table 2. Sample of the weights (in *g* of newborn babies in 1978

3840	3540	3920	2920	3820	3910	3300	2770	3000	3900
3150	3180	3160	3040	3920	2900	3420	2780	3500	3310
3760	3280	3720	2480	3320	3200	3200	2850	3340	3460
3020	2830	2900	3400	4020	3360	3300	3200	2500	2060
3150	3640	3380	3400	3220	2790	2420	3390	3680	3160
3220	3200	2790	3000	3070	2820	2880	3480	2880	2880

skewness. It is defined as

$$\text{skewness} = \frac{\text{mean} - \text{median}}{s}.$$

The values are 0.053 for the data in Table 1 (1988), and 0.039 for the data in Table 2 (1978). Both are positive. According to this new skewness, the right tail is longer than the left one in both distributions. Each mean is much more affected by the right tail than the median.

Using this definition, the mean is greater than the median, if and only if skewness is positive. Under this new skewness there is no counter-example to principle 2.

Ⅴ. Conclusion

The principles stated above are important and true in general cases. However, beginners tend to forget counter-examples. The purpose of this paper is to show an example and a counter-example to a principle. It is important for teachers to show beginners the existence of a counter-example.

Acknowledgment

The author thanks S. Ishimura for providing an interesting sample [3].

References

[1] 中澤　港、「R による統計解析の基礎」、ピアソン・エデュケーション、2005。

[2] 石村貞夫、デズモンド・アレン、「すぐわかる統計用語」、東京図書、2006。

[3] 石村貞夫、「すぐわかる統計解析」、東京図書、2007。

[4] Gerald van Belle, “Statistical Rules of Thumb,” Wiley, 2002.

[5] Crawley, M. J. “Statistics－An Intro-duction using R,” Wiley, 2007.

[6] R Development Core Team, “R: A Language and Environment for Statistical Computing,” R Foundation for Statistical Computing, 2009.

Appendix

```
stdev <-function(x) {
  sqrt(
    sum((x-mean(x))^2)/length(x)
  )
}
skew <-function(x) {
  m3 <-sum((x-mean(x))^3/length(x)
  s3 <-stdev(x)^3
  m3/s3
}
```